R E P U B L I Q U E F R A N Ç A I S E

10/779 858
June 09 2004
20

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 10 MARS 2004

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS cedex 08
Téléphone : 33 (0)1 53 04 53 04
Télécopie : 33 (0)1 53 04 45 23
www.inpi.fr

ETABLISSEMENT PUBLIC NATIONAL CREE PAR LA LOI N° 51-444 DU 19 AVRIL 1951

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ
Code de la propriété intellectuelle - Livre VI

cerfa N° 11354*01

REQUÊTE EN DÉLIVRANCE 1/2

Cet imprimé est à remplir lisiblement à l'encre noire

DB 540 W /260899

Réservé à l'INPI

REMISE DES PIÈCES
DATE 14 FEV 2003
LIEU 75 INPI PARIS
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0301812
DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI 14 FEV. 2003

Vos références pour ce dossier *(facultatif)* BR 7747/VR/MB

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

Cabinet LHERMET LA BIGNE & REMY
191, rue Saint-Honoré
75001 PARIS
France

Confirmation d'un dépôt par télécopie ☐ N° attribué par l'INPI à la télécopie

2 NATURE DE LA DEMANDE	Cochez l'une des 4 cases suivantes	
Demande de brevet	☒	
Demande de certificat d'utilité	☐	
Demande divisionnaire	☐	
Demande de brevet initiale	N°	Date / /
ou demande de certificat d'utilité initiale	N°	Date / /
Transformation d'une demande de brevet européen *Demande de brevet initiale*	☐ N°	Date / /

3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Procédé de classification hiérarchique descendante de données multi-valuées

4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE	
	Pays ou organisation Date / / N°
	Pays ou organisation Date / / N°
	Pays ou organisation Date / / N°
	☐ S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»

5 DEMANDEUR		☐ S'il y a d'autres demandeurs, cochez la case et utilisez l'imprimé «Suite»
Nom ou dénomination sociale		FRANCE TELECOM
Prénoms		
Forme juridique		
N° SIREN		
Code APE-NAF		
Adresse	Rue	6 place d'Alleray
	Code postal et ville	75015 PARIS
Pays		
Nationalité		
N° de téléphone *(facultatif)*		
N° de télécopie *(facultatif)*		
Adresse électronique *(facultatif)*		

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE 2/2

Réservé à l'INPI

REMISE DES PIÈCES
DATE 14 FEV 2003
LIEU 75 INPI PARIS
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0301812

DB 540 W /260899

Vos références pour ce dossier : *(facultatif)* BR 7747/VR/MB		
6 MANDATAIRE		
Nom		
Prénom		
Cabinet ou Société		Cabinet LHERMET LA BIGNE & REMY
N °de pouvoir permanent et/ou de lien contractuel		
Adresse	Rue	191, rue Saint-Honoré
	Code postal et ville	75001 PARIS
N° de téléphone *(facultatif)*		01 44 77 80 00
N° de télécopie *(facultatif)*		01 44 77 88 44
Adresse électronique *(facultatif)*		cabinet@lhermetlabigneremy.fr
7 INVENTEUR (S)		
Les inventeurs sont les demandeurs		☐ Oui ☒ Non **Dans ce cas fournir une désignation d'inventeur(s) séparée**
8 RAPPORT DE RECHERCHE		**Uniquement pour une demande de brevet (y compris division et transformation)**
Établissement immédiat ou établissement différé		☒ ☐
Paiement échelonné de la redevance		**Paiement en deux versements, uniquement pour les personnes physiques** ☐ Oui ☒ Non
9 RÉDUCTION DU TAUX DES REDEVANCES		**Uniquement pour les personnes physiques** ☐ Requise pour la première fois pour cette invention *(joindre un avis de non-imposition)* ☐ Requise antérieurement à ce dépôt *(joindre une copie de la décision d'admission pour cette invention ou indiquer sa référence) :*
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes		

10 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (Nom et qualité du signataire)

Cabinet LHERMET LA BIGNE & REMY

Vincent REMY (CPI n° 96-0701)

VISA DE LA PRÉFECTURE OU DE L'INPI

L. MARIELLO

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire. Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

La présente invention concerne un procédé de classification hiérarchique descendante de données, chaque donnée étant associée à des valeurs particulières initiales d'attributs communs aux données. Plus particulièrement, l'invention concerne un procédé de classification comprenant des étapes récursives de divisions d'ensembles de données.

Le procédé de classification automatique de Williams & Lambert est un procédé de ce type. Il s'applique cependant à des données dont les attributs sont binaires, c'est-à-dire des attributs prenant pour chaque donnée une valeur particulière « Vrai » ou « Faux ». Selon ce procédé, lors de chaque étape de division d'un ensemble, on calcule pour chaque attribut la valeur du Khi2 cumulé sur tous les autres attributs (la valeur du Khi2 calculé entre deux attributs permet d'estimer le lien entre ces deux attributs). On divise ensuite l'ensemble en sous-ensembles sur la base de l'attribut ayant la valeur du Khi2 cumulé la plus élevée.

Ce procédé peut être étendu à la classification de données dont les attributs prennent des valeurs symboliques, moyennant l'exécution d'une étape préliminaire dite de "binarisation". Lors de cette étape chaque valeur symbolique qu'un attribut peut prendre est transformée en un attribut binaire. Ensuite, au cours des étapes récursives de division, on calcule les valeurs du Khi2 sur les matrices de contingence des couples d'attributs binaires obtenus.

Cependant, ce procédé ne peut pas être appliqué sans inconvénient majeur à la classification de données multi-valuées mixtes numériques/symboliques, c'est-à-dire des données dont certains attributs sont symboliques et d'autres numériques Dans ce document, nous entendons par valeurs numériques des valeurs quantitatives (représentées par des nombres) et par valeurs symboliques des valeurs qualitatives (dites aussi discrètes, et représentables par exemple par des lettres ou des mots").

En effet, en ce qui concerne les attributs numériques, une discrétisation préliminaire des valeurs par intervalles est nécessaire, de manière à rendre symbolique chaque attribut numérique. Or cette transformation fait inévitablement perdre de l'information, sans compter que le nombre d'intervalles de discrétisation va influer sur le résultat final, sans qu'il soit possible de choisir judicieusement ce nombre d'intervalles a priori. La cohérence des classes obtenues s'en trouve affectée.

De plus, même dans le cas d'attributs uniquement symboliques, l'étape préliminaire de "binarisation" augmente considérablement le nombre d'attributs, ce qui augmente également considérablement le temps d'exécution du procédé.

Enfin, le calcul du Khi2 est une estimation du lien entre deux attributs, et met en valeur des attributs corrélés ou anti-corrélés. Ce calcul surestime donc artificiellement le

lien entre des attributs anti-corrélés issus de l'étape de binarisation. Le calcul du Khi2 étant en outre symétrique entre deux variables, il ne permet pas de déterminer si une variable est plus discriminante qu'une autre.

L'invention vise à remédier à ces inconvénients en fournissant un procédé de classification hiérarchique descendante capable de traiter des données multi-valuées numériques et/ou symboliques en optimisant la complexité de traitement et la cohérence des classes obtenues.

L'invention a donc pour objet un procédé de classification hiérarchique descendante de données, chaque donnée étant associée à des valeurs particulières initiales d'attributs communs aux données, le procédé comprenant des étapes récursives de divisions d'ensembles de données, caractérisé en ce que, lors de chaque étape de division d'un ensemble, on calcule des valeurs discrètes d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et en ce que l'on divise ledit ensemble en sous-ensembles en fonction des valeurs discrètes.

En effet, lors de l'exécution d'un procédé de classification selon l'invention, on calcule de nouvelles valeurs discrètes d'attributs associées à des données que l'on souhaite classer, à chaque étape récursive de division du procédé. Cette discrétisation n'étant pas réalisée une bonne fois pour toute lors d'une étape préliminaire, aucune information n'est perdue lors de l'exécution du procédé. De plus, à chaque itération, la division d'un ensemble en sous-ensembles se basant sur les valeurs discrètes des attributs calculés temporairement, le procédé en est d'autant simplifié.

De façon optionnelle, lors de chaque étape de division d'un ensemble, on calcule des valeurs binaires d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et l'on divise ledit ensemble en sous-ensembles en fonction des valeurs binaires.

Ce principe de discrétisation de chaque attribut numérique et symbolique en seulement deux valeurs (dit "binarisation", de l'anglais "binning") maximise la vitesse d'exécution de l'algorithme sans nuire sensiblement à sa précision sur de grands volumes de données.

Un procédé de classification selon l'invention peut en outre comporter l'une ou plusieurs des caractéristiques suivantes :

- lors de l'étape de calcul des valeurs binaires d'attributs, on calcule pour chaque attribut numérique une estimation de la médiane des valeurs particulières initiales de cet attribut pour les données dudit ensemble, et l'on affecte à l'attribut binaire correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la

valeur particulière initiale de l'attribut numérique pour cette donnée est inférieure ou égale à l'estimation de la médiane, et la valeur « Faux » sinon ;

- l'estimation de la médiane d'un attribut numérique est obtenue de la façon suivante :

  - on extrait des valeurs extrêmes de l'ensemble des valeurs prises par l'attribut numérique pour les données dudit ensemble ;
  - on calcule la moyenne des valeurs restantes ; et
  - on affecte à l'estimation de la médiane la valeur de cette moyenne.

- lors de l'étape de calcul des valeurs binaires d'attributs, on calcule pour chaque attribut symbolique une estimation du mode des valeurs particulières initiales de cet attribut pour les données dudit ensemble, et l'on affecte à l'attribut binaire correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la valeur particulière initiale de l'attribut numérique pour cette donnée est égale à l'estimation du mode, et la valeur « Faux » sinon ;

- l'estimation du mode d'un attribut symbolique est obtenue de la façon suivante :

  - on mémorise les m premières valeurs symboliques différentes prises par les données dudit ensemble pour l'attribut symbolique, m étant un nombre prédéterminé ;
  - on retient la valeur symbolique apparaissant le plus souvent parmi ces m premières valeurs symboliques différentes ; et
  - on affecte à l'estimation du mode cette valeur symbolique retenue.

- on divise ledit ensemble en sous-ensembles en fonction d'un critère d'homogénéité calculé à partir des valeurs discrètes d'attributs dudit ensemble ;

- on divise ledit ensemble sur la base des valeurs discrètes de l'attribut le plus discriminant, c'est à dire l'attribut pour lequel un critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus est optimisé ;

- pour un attribut quelconque le critère d'homogénéité est une estimation de l'espérance des probabilités conditionnelles de prédire correctement les autres attributs connaissant cet attribut ; et

- certains attributs étant a priori marqués comme tabous au moyen d'un paramètre particulier, l'attribut le plus discriminant est l'attribut non marqué tabou pour lequel le critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus est optimisé.

L'invention sera mieux comprise à l'aide de la description qui va suivre, donnée uniquement à titre d'exemple et faite en se référant aux dessins annexés dans lesquels :

- la figure 1 illustre schématiquement la structure d'un système informatique pour la mise en œuvre d'un procédé selon l'invention, ainsi que la structure de données fournies en entrée et en sortie de ce système ; et
- la figure 2 représente les étapes successives d'un procédé selon l'invention.

Le système représenté sur la figure 1 est un système informatique classique comprenant un calculateur 10 associé à des mémoires de type RAM et ROM (non représentées) pour le stockage de données 12 et 14 fournies en entrée et en sortie du calculateur 10. Les données 12 fournies en entrée du calculateur 10 sont par exemple stockées sous la forme d'une base de données, ou bien sous la forme d'un simple fichier. Les données fournies en sortie du calculateur 10 sont stockées dans un format qui permet, pour la mise en œuvre du procédé selon l'invention, de les représenter sous la forme d'une structure arborescente, telle qu'un arbre de décision 14.

Les données 12 sont des données multi-valuées numériques et/ou symboliques. Ces données sont par exemple issues de bases de données médicales, marketing, c'est-à-dire des bases de données contenant généralement plusieurs millions de données associées chacune à plusieurs dizaines d'attributs numériques ou symboliques.

Dans la suite de la description, l'ensemble des données sera noté $D = \{d_1, \ldots, d_n\}$. L'ensemble des attributs sera noté $A = \{a_1, \ldots, a_p\}$. Ainsi, chaque donnée d_i multi-valuée peut être représentée dans l'espace A des attributs, sous la forme suivante :

$d_i = (a_1 (d_i) ; \ldots ; a_p (d_i))$, où $a_j (d_i)$ est la valeur que prend l'attribut a_j pour la donnée d_i.

Les attributs a_j peuvent être numériques ou symboliques. Par exemple, comme représenté sur la figure 1, l'attribut a_1 est numérique. Il prend la valeur 12 pour la donnée d_1 et la valeur 95 pour la donnée d_n. L'attribut a_p est symbolique. Il attribue par exemple une couleur aux données de la base : ainsi la donnée d_1 est de couleur bleue et la donnée d_n est de couleur rouge.

Il est judicieux de représenter cette base de données multi-valuées sous la forme d'un tableau dont les lignes correspondent chacune à une donnée d_i et dont les colonnes correspondent chacune à un attribut a_j.

Le calculateur 10 met en œuvre un procédé de classification automatique hiérarchique descendante de ces données 12 multi-valuées numériques et/ou symboliques, dont l'objectif est de générer des classes homogènes de ces données, classes auxquelles on accède à l'aide de l'arbre de décision 14 associé.

Un mode de réalisation préféré de l'invention est d'organiser les classes obtenues en un arbre de décision binaire, c'est-à-dire un mode de réalisation dans lequel on divise

une classe de données en deux sous-classes. Ce mode de réalisation particulièrement simple permet une classification rapide et efficace des données.

Pour la mise en œuvre du procédé de classification, le calculateur 10 comporte un module pilote 16 dont la fonction est de coordonner l'activation d'un module d'entrées/sorties 18, d'un module de discrétisation 20 et d'un module de segmentation 22. En synchronisant ces trois modules, il permet la génération récursive de l'arbre de décision 14 et des classes homogènes.

Le module d'entrées/sorties 18 a pour fonction de lire les données 12 fournies en entrée du calculateur 10. En particulier, il a pour fonction d'identifier le nombre de données à traiter et le type des attributs associés à ces données, pour les fournir au module de discrétisation 20.

Le module de discrétisation 20 a pour fonction de transformer les attributs $a_1, \ldots, a_p$ en attributs discrets. Plus précisément, dans cet exemple, le module de discrétisation 20 est un module de binarisation qui a pour fonction de transformer chaque attribut en attribut binaire, c'est-à-dire en attribut pouvant uniquement prendre la valeur Vrai ou Faux pour chacune des données d_i. Son fonctionnement sera détaillé en référence à la figure 2.

Le module de segmentation 22 a pour fonction de déterminer, parmi les attributs binaires calculées par le module de binarisation 20, celui qui est le plus discriminant pour diviser un ensemble de données en deux sous-ensembles les plus homogènes possibles. Son fonctionnement sera détaillé en référence à la figure 2.

Le procédé récursif de classification automatique et de génération d'un arbre de décision associé comporte une première étape 30 d'extraction de données de la base de données 12. Lors de cette étape, il s'agit d'extraire de la base 12 les données appartenant à un ensemble E_1, représenté par un nœud terminal de l'arbre de décision 14, et que l'on souhaite diviser en deux sous-ensembles E_{11} et E_{12}.

Ces données sont extraites avec leurs attributs et ceux-ci sont fournis en entrée du module de binarisation 20, qui traite séparément les attributs symboliques et les attributs numériques.

Ainsi, lors d'une étape 32a d'estimation de valeur médiane, le module de binarisation 20 calcule, pour chaque attribut numérique a_j, une estimation de la valeur médiane de l'ensemble des valeurs suivantes :

$\{d_1 (a_j) ; \ldots ; d_n (a_j)\}$.

Lors de cette étape 32a, il est possible de calculer directement la valeur médiane M_j de l'ensemble des valeurs prises par l'attribut a_j, mais ce calcul peut être remplacé par un procédé d'estimation de cette valeur médiane, plus simple à mettre en œuvre par des moyens informatiques.

Ce procédé d'estimation de la médiane M_j comporte par exemple les étapes suivantes :

- on extrait des valeurs extrêmes de l'ensemble des valeurs prises par l'attribut a_j ;
- on calcule la moyenne des valeurs restantes ; et
- on affecte à M_j la valeur de cette moyenne.

Les valeurs extrêmes extraites de l'ensemble sont par exemple, n valeurs maximales et n valeurs minimales, n étant un paramètre prédéterminé ou résultant d'une analyse préalable de la distribution des valeurs prises par l'attribut a_j.

Il est également possible d'estimer la valeur de la médiane par le simple calcul de la moyenne de l'ensemble des valeurs de l'attribut.

Lors de l'étape suivante 34a de calcul d'attributs binaires, on calcule les valeurs d'un attribut binaire b_j, à partir de chaque attribut numérique a_j, de la façon suivante :

$$\text{si } d_i\left(a_j\right) \leq M_j, d_i\left(b_j\right) = \text{vrai};$$

$$\text{si } d_i\left(a_j\right) > M_j, d_i\left(b_j\right) = \text{faux}.$$

En ce qui concerne les attributs symboliques a_k, le module de binarisation 20 calcule, pour chacun d'entre eux, une estimation du mode de leurs valeurs. Ceci est réalisé lors d'une étape 32b d'estimation de mode.

Le mode M_k d'un ensemble de valeurs symboliques d'un attribut a_k est la valeur symbolique prise le plus souvent par cet attribut.

Ce mode M_k peut être calculé mais cela est coûteux en temps de calcul.

Pour simplifier cette étape, on peut remplacer le calcul direct du mode par un procédé d'estimation de celui-ci comportant les étapes suivantes :

- lors de la lecture les données de l'ensemble E1, le module de binarisation 20 mémorise les m premières valeurs symboliques différentes prises par les données di pour l'attribut a_k, m étant un nombre prédéterminé ;
- on retient la valeur symbolique apparaissant le plus souvent parmi ces m premières valeurs symboliques différentes ; et
- on affecte cette valeur symbolique retenue au mode M_k.

On choisit par exemple m = 200.

Si l'attribut a_k comporte un nombre de valeurs symboliques possibles inférieur à m, alors l'estimation du mode M_k est égale au mode lui-même. Sinon, l'estimation du mode M_k a de fortes chances de constituer une bonne valeur de remplacement du mode dans

de nombreux cas. D'une façon générale, la plupart des attributs statistiques symboliques ont moins de quelques dizaines de valeurs symboliques différentes.

Lors de l'étape 34b suivante de calcul d'attributs binaires, on calcule les valeurs d'un attribut binaire b_k, à partir de chaque attribut symbolique a_k, de la façon suivante :

$$\text{si } d_i(a_k) = M_k, d_i(b_k) = \text{vrai};$$
$$\text{si } d_i(a_k) \neq M_k, d_i(b_k) = \text{faux}.$$

Suite aux étapes 34a et 34b, on passe à une étape 36 lors de laquelle on rassemble les attributs binaires b_k, b_j issus des attributs symboliques a_k et numériques a_j. On constitue ainsi un ensemble $B = \{b_1,\ldots,b_p\}$ d'attributs binaires pour l'ensemble E_1 des données d_i. Lors de cette étape, le module de binarisation 20 fournit les données multi-valuées de l'ensemble E_1 associées à leurs attributs binaires $\{b_1\ldots, b_p\}$ au module de segmentation 22.

Ensuite, lors d'une étape de calcul 38, le module de segmentation 22 calcule pour chaque attribut b_j la valeur f (b_j) suivante :

$$f(b_j) = \sum_{k,k\neq j} FU(b_j, b_k), \text{ avec}$$

$$FU(b_j, b_k) = \frac{1}{n}\left[\begin{array}{l} c(B_j)\,\mathrm{Max}\left(p(B_k / B_j); p(\neg B_k / B_j)\right) + \\ c(\neg B_j)\,\mathrm{Max}\left(p(B_k / \neg B_j); p(\neg B_k / \neg B_j)\right) \end{array}\right]$$

où pour tout indice j, B_j est l'événement « l'attribut b_j prend la valeur Vrai » ; et

$\neg B_j$ est l'événement « l'attribut b_j prend la valeur Faux »,

avec Max(x,y) : fonction retournant le maximum entre x et y ;

p(x/y) : probabilité de l'événement x sachant l'événement y ; et

c(x) effectif de l'événement x (pondération).

Telle qu'elle est présentée ci-dessus, pour chaque attribut b_j, la valeur $f(b_j)$ est une estimation de l'espérance des probabilités conditionnelles de prédire correctement les autres attributs, connaissant la valeur de l'attribut b_j. En d'autres termes, elle permet d'évaluer la pertinence d'une segmentation en deux sous-ensembles basée sur l'attribut b_j.

Une autre fonction f peut cependant être choisie pour optimiser la segmentation, telle qu'une fonction basée sur un calcul de covariance des attributs.

Lors de l'étape de sélection 40 suivante, le module de segmentation 22 détermine l'attribut binaire b_{jmax} qui maximise la valeur $f(b_{jmax})$, c'est à dire l'attribut le plus discriminant pour une segmentation en deux sous-ensembles.

Ensuite, lors d'une étape 42 de segmentation, le module 22 génère deux sous-ensembles E_{11} et E_{12} à partir de l'ensemble des données E_1. Le premier ensemble E_{11} est

par exemple le sous-ensemble regroupant les données pour lesquelles l'attribut b_{jmax} prend la valeur Vrai et le sous-ensemble E_{12} regroupe les données de l'ensemble E_1 pour lesquelles l'attribut b_{jmax} prend la valeur Faux.

Lors de cette étape, on met à jour l'arbre de décision 14 en rajoutant deux nœuds E_{11} et E_{12} reliés au nœud E_1 par deux nouvelles branches.

Ainsi, lorsque l'on se déplace dans cet arbre de décision et que l'on arrive au nœud E_1, on effectue le test suivant :

"la donnée d_i a t'elle, pour l'attribut a_{jmax}, une valeur inférieure à M_{jmax}?", si a_{jmax} est un attribut numérique ; ou

"la donnée d_i a t'elle, pour l'attribut a_j Max, une valeur égale à M_{jmax} ?", si a_{jmax} est un attribut symbolique.

Si la réponse à ce test est positive, alors la donnée d_i appartient au sous-ensemble E_{11}, sinon elle appartient au sous-ensemble E_{12}.

Suite à l'étape 42, lors d'une étape 44 de test, on teste un critère d'arrêt du procédé. Ce critère d'arrêt est par exemple le nombre de nœuds terminaux de l'arbre de décision, c'est-à-dire le nombre de classes obtenues par le procédé de classification, si l'on s'est fixé un nombre de classes à ne pas dépasser.

Le critère d'arrêt peut aussi être le nombre de niveaux dans l'arbre de décision. On peut également imaginer d'autres critères d'arrêt.

Si ce critère d'arrêt est atteint, on passe à une étape 46 de fin de procédé. Sinon on passe à l'étape 30 lors de laquelle on recommence le procédé décrit précédemment à partir d'un nouvel ensemble de données, par exemple l'ensemble E_{11} ou l'ensemble E_{12} obtenu précédemment.

On notera que le procédé de classification décrit précédemment est un procédé non supervisé.

Ce procédé de classification peut également être utilisé en mode "semi-supervisé". L'application d'un procédé de classification en mode semi-supervisé est utile lorsque l'on souhaite prédire ou expliquer un attribut particulier en fonction de tous les autres alors que cet attribut particulier est mal ou peu renseigné dans la base de données 12, c'est-à-dire lorsque pour un grand nombre de données d_i, aucune valeur ne correspond à cet attribut. Il suffit dans ce cas d'identifier cet attribut comme purement "à expliquer", et de le marquer comme tel via un marquage particulier, par exemple dans un fichier de paramètres associés. Cet attribut spécifié comme "à expliquer" par l'utilisateur est dit attribut "tabou". L'attribut tabou ne doit pas être choisi comme discriminant.

On notera aussi que l'on peut définir plusieurs attributs tabous. Il suffit dans ce cas de distinguer parmi les attributs a_j, les attributs dits "explicatifs" et les attributs "tabous".

On s'interdit alors de sélectionner les attributs tabous comme attributs discriminants pour effectuer une segmentation, lors de l'étape 40 précédemment décrite.

En effet, en mode semi-supervisé, lors de l'étape 40, si l'attribut sélectionné est un attribut tabou, alors on cherche le deuxième attribut qui maximise la fonction $f(b_j)$ et ainsi de suite jusqu'à trouver l'attribut non tabou le plus discriminant, c'est-à-dire celui qui maximise le critère d'homogénéité des valeurs discrétisées des autres attributs dans les sous-ensembles E_{11} et E_{12}.

La classification finalement obtenue permettra ensuite de prédire les valeurs d'un attribut tabou, pour les données où celles-ci sont manquantes. En effet, le procédé de classification effectue des tests uniquement sur l'ensemble des attributs explicatifs tout en exploitant au maximum toutes les corrélations entre attributs.

La prédiction des valeurs d'un attribut tabou se fait en remplaçant des valeurs manquantes ou mal renseignées par les valeurs renseignées les plus probables dans chaque classe.

Il apparaît clairement qu'un procédé selon l'invention permet la classification simple et efficace selon un mode hiérarchique descendant, de données multi-valuées numériques et/ou symboliques. Sa faible complexité permet de l'envisager pour la classification de grandes bases de données.

REVENDICATIONS

1. Procédé de classification hiérarchique descendante de données (12), chaque donnée (12) étant associée à des valeurs particulières initiales d'attributs ($a_1,\ldots,a_p$) communs aux données, le procédé comprenant des étapes récursives (32a, 32b, 34a, 34b, 36, 38, 40, 42) de divisions d'ensembles (E_1, E_{11}, E_{12}) de données, caractérisé en ce que, lors de chaque étape de division d'un ensemble (E_1), on calcule (32a, 32b, 34a, 34b, 36) des valeurs discrètes d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et en ce que l'on divise (38, 40, 42), ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction des valeurs discrètes.

2. Procédé de classification hiérarchique descendante de données (12) selon la revendication 1, caractérisé en ce que, lors de chaque étape de division d'un ensemble (E_1), on calcule (32a, 32b, 34a, 34b, 36) des valeurs binaires d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et en ce que l'on divise (38, 40, 42) ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction des valeurs binaires.

3. Procédé de classification hiérarchique descendante de données (12) selon la revendication 1 ou 2, caractérisé en ce que lors de l'étape (32a, 32b, 34a, 34b, 36) de calcul des valeurs binaires d'attributs, on calcule (32a) pour chaque attribut numérique une estimation de la médiane des valeurs particulières initiales de cet attribut pour les données dudit ensemble, et en ce que l'on affecte (34a) à l'attribut binaire correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la valeur particulière initiale de l'attribut numérique pour cette donnée est inférieure ou égale à l'estimation de la médiane, et la valeur « Faux » sinon.

4. Procédé de classification hiérarchique descendante de données (12) selon la revendication 3, caractérisé en ce que l'estimation de la médiane d'un attribut numérique est obtenue de la façon suivante :

- on extrait des valeurs extrêmes de l'ensemble des valeurs prises par l'attribut numérique pour les données dudit ensemble ;
- on calcule la moyenne des valeurs restantes ; et
- on affecte à l'estimation de la médiane la valeur de cette moyenne.

5. Procédé de classification hiérarchique descendante de données (12) selon l'une quelconque des revendications 1 à 4, caractérisé en ce que lors de l'étape (32a, 32b, 34a, 34b, 36) de calcul des valeurs binaires d'attributs, on calcule (32b) pour chaque attribut symbolique une estimation du mode des valeurs particulières initiales de cet attribut pour les données dudit ensemble, et en ce que l'on affecte (34b) à l'attribut binaire

REVENDICATIONS

1. Procédé de classification hiérarchique descendante de données multi-valuées (12) stockées dans des moyens de stockage d'un système informatique, chaque donnée (12) étant associée à des valeurs particulières initiales d'attributs ($a_1,...,a_p$) communs aux données, le procédé comprenant des étapes récursives (32a, 32b, 34a, 34b, 36, 38, 40, 42) de divisions d'ensembles (E_1, E_{11}, E_{12}) de données, caractérisé en ce que, lors de chaque étape de division d'un ensemble (E_1), on calcule (32a, 32b, 34a, 34b, 36) des valeurs discrètes d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et en ce que l'on divise (38, 40, 42) ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction des valeurs discrètes.

2. Procédé de classification hiérarchique descendante de données (12) selon la revendication 1, caractérisé en ce que, lors de chaque étape de division d'un ensemble (E_1), on calcule (32a, 32b, 34a, 34b, 36) des valeurs binaires d'attributs à partir des valeurs particulières initiales d'attributs des données dudit ensemble, et en ce que l'on divise (38, 40, 42) ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction des valeurs binaires.

3. Procédé de classification hiérarchique descendante de données (12) selon la revendication 1 ou 2, caractérisé en ce que lors de l'étape (32a, 32b, 34a, 34b, 36) de calcul des valeurs binaires d'attributs, on calcule (32a) pour chaque attribut numérique une estimation de la médiane des valeurs particulières initiales de cet attribut pour les données dudit ensemble, et en ce que l'on affecte (34a) à l'attribut binaire correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la valeur particulière initiale de l'attribut numérique pour cette donnée est inférieure ou égale à l'estimation de la médiane, et la valeur « Faux » sinon.

4. Procédé de classification hiérarchique descendante de données (12) selon la revendication 3, caractérisé en ce que l'estimation de la médiane d'un attribut numérique est obtenue de la façon suivante :

- on extrait des valeurs extrêmes de l'ensemble des valeurs prises par l'attribut numérique pour les données dudit ensemble ;
- on calcule la moyenne des valeurs restantes ; et
- on affecte à l'estimation de la médiane la valeur de cette moyenne.

5. Procédé de classification hiérarchique descendante de données (12) selon l'une quelconque des revendications 1 à 4, caractérisé en ce que lors de l'étape (32a, 32b, 34a, 34b, 36) de calcul des valeurs binaires d'attributs, on calcule (32b) pour chaque attribut symbolique une estimation du mode des valeurs particulières initiales de cet attribut pour

correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la valeur particulière initiale de l'attribut numérique pour cette donnée est égale à l'estimation du mode, et la valeur « Faux » sinon.

6. Procédé de classification hiérarchique descendante de données (12) selon la revendication 5, caractérisé en ce que l'estimation du mode d'un attribut symbolique est obtenue de la façon suivante :

- on mémorise les m premières valeurs symboliques différentes prises par les données dudit ensemble pour l'attribut symbolique, m étant un nombre prédéterminé ;
- on retient la valeur symbolique apparaissant le plus souvent parmi ces m premières valeurs symboliques différentes ; et
- on affecte à l'estimation du mode cette valeur symbolique retenue.

7. Procédé de classification selon l'une quelconque des revendications 1 à 6, caractérisé en ce que l'on divise ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction d'un critère d'homogénéité calculé à partir des valeurs discrètes d'attributs dudit ensemble (E_1).

8. Procédé de classification selon l'une quelconque des revendications 1 à 7, caractérisé en ce que l'on divise ledit ensemble (E_1) sur la base des valeurs discrètes de l'attribut le plus discriminant, c'est à dire l'attribut pour lequel un critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus (E_{11}, E_{12}) est optimisé.

9. Procédé de classification selon la revendication 8, caractérisé en ce que pour un attribut quelconque le critère d'homogénéité est une estimation de l'espérance des probabilités conditionnelles de prédire correctement les autres attributs connaissant cet attribut.

10. Procédé de classification selon la revendication 8 ou 9, caractérisé en ce que, certains attributs étant a priori marqués comme tabous au moyen d'un paramètre particulier, l'attribut le plus discriminant est l'attribut non marqué comme tabou pour lequel le critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus (E_{11}, E_{12}) est optimisé.

les données dudit ensemble, et en ce que l'on affecte (34b) à l'attribut binaire correspondant à cet attribut pour une donnée dudit ensemble, la valeur « Vrai » si la valeur particulière initiale de l'attribut numérique pour cette donnée est égale à l'estimation du mode, et la valeur « Faux » sinon.

6. Procédé de classification hiérarchique descendante de données (12) selon la revendication 5, caractérisé en ce que l'estimation du mode d'un attribut symbolique est obtenue de la façon suivante :

- on mémorise les m premières valeurs symboliques différentes prises par les données dudit ensemble pour l'attribut symbolique, m étant un nombre prédéterminé ;
- on retient la valeur symbolique apparaissant le plus souvent parmi ces m premières valeurs symboliques différentes ; et
- on affecte à l'estimation du mode cette valeur symbolique retenue.

7. Procédé de classification selon l'une quelconque des revendications 1 à 6, caractérisé en ce que l'on divise ledit ensemble (E_1) en sous-ensembles (E_{11}, E_{12}) en fonction d'un critère d'homogénéité calculé à partir des valeurs discrètes d'attributs dudit ensemble (E_1).

8. Procédé de classification selon l'une quelconque des revendications 1 à 7, caractérisé en ce que l'on divise ledit ensemble (E_1) sur la base des valeurs discrètes de l'attribut le plus discriminant, c'est à dire l'attribut pour lequel un critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus (E_{11}, E_{12}) est optimisé.

9. Procédé de classification selon la revendication 8, caractérisé en ce que pour un attribut quelconque le critère d'homogénéité est une estimation de l'espérance des probabilités conditionnelles de prédire correctement les autres attributs connaissant cet attribut.

10. Procédé de classification selon la revendication 8 ou 9, caractérisé en ce que, certains attributs étant a priori marqués comme tabous au moyen d'un paramètre particulier, l'attribut le plus discriminant est l'attribut non marqué comme tabou pour lequel le critère d'homogénéité de l'ensemble des valeurs discrètes des autres attributs dans les sous-ensembles obtenus (E_{11}, E_{12}) est optimisé.

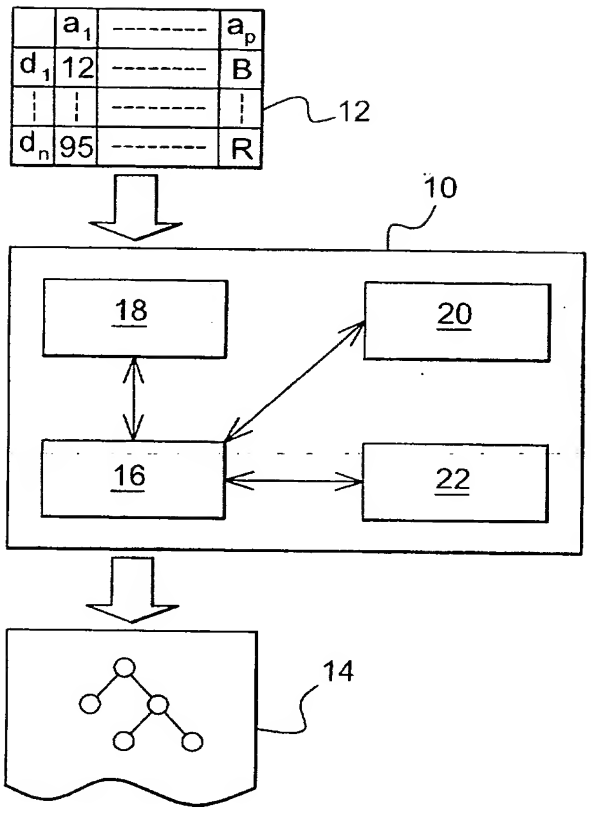

Fig. 1

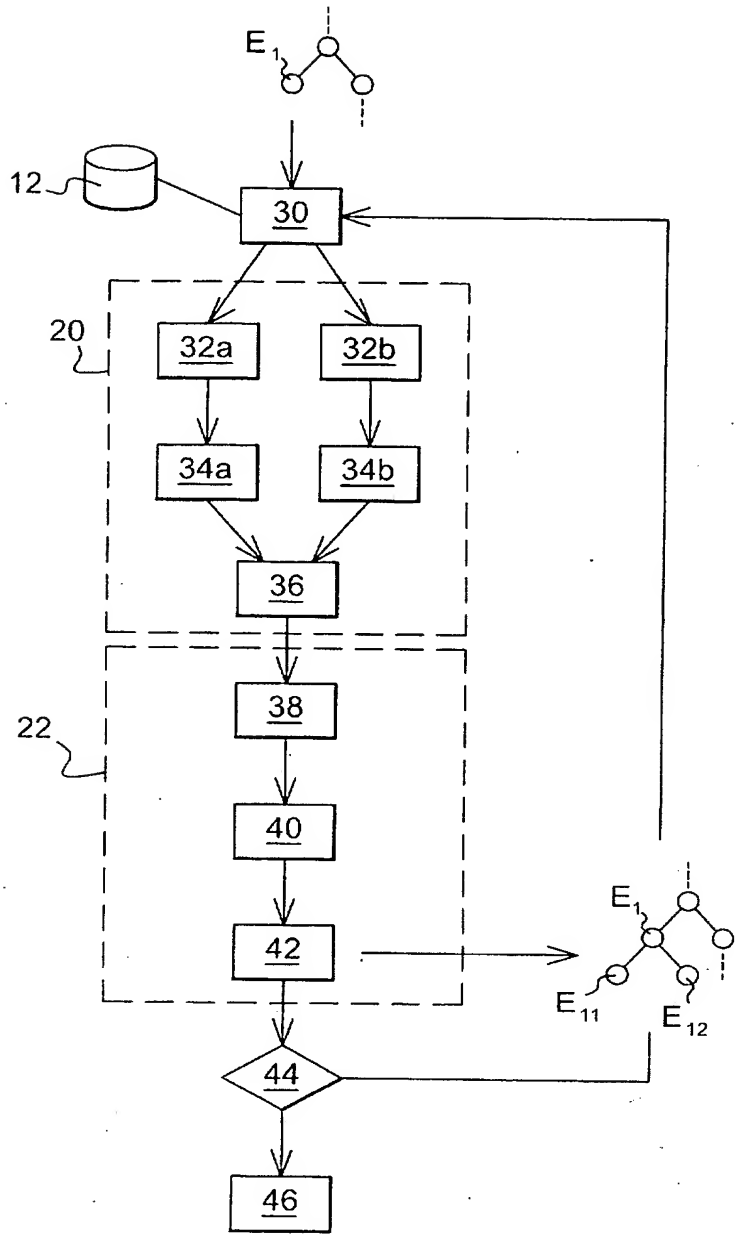

Fig. 2

reçue le 26/03/03

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

DÉPARTEMENT DES BREVETS

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 01 53 04 53 04 Télécopie : 01 42 93 59 30

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11 235*02

DÉSIGNATION D'INVENTEUR(S) Page N° 1 / 1

(Si le demandeur n'est pas l'inventeur ou l'unique inventeur)

Cet imprimé est à remplir lisiblement à l'encre noire

DB 113 W /260899

Vos références pour ce dossier *(facultatif)*	BR 7747/VR/MB
N° D'ENREGISTREMENT NATIONAL	0301812

TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Procédé de classification hiérarchique descendante de données multi-valuées

LE(S) DEMANDEUR(S) :

FRANCE TELECOM
6 place d'Alleray
F - 75015 PARIS

DESIGNE(NT) EN TANT QU'INVENTEUR(S) : (Indiquez en haut à droite «Page N° 1/1» S'il y a plus de trois inventeurs, utilisez un formulaire identique et numérotez chaque page en indiquant le nombre total de pages).

Nom		MEYER
Prénoms		Frank
Adresse	Rue	1A, rue Sully Prud'homme
	Code postal et ville	22300 LANNION
Société d'appartenance *(facultatif)*		
Nom		
Prénoms		
Adresse	Rue	
	Code postal et ville	
Société d'appartenance *(facultatif)*		
Nom		
Prénoms		
Adresse	Rue	
	Code postal et ville	
Société d'appartenance *(facultatif)*		

DATE ET SIGNATURE(S)
DU (DES) DEMANDEUR(S)
OU DU MANDATAIRE
(Nom et qualité du signataire)

Paris, le 13 février 2003
Vincent REMY (CPI n°96/0701)

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire.
Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.